# Beyond Digital "Echo Chambers": The Role of Viewpoint Diversity in Political Discussion

Rishav Hada
rishavhada@gmail.com
Microsoft Research India
Bengaluru, India

Amir Ebrahimi Fard
a.ebrahimifard@maastrichtuniversity.nl
Maastricht University
Maastricht, Netherlands

Sarah Shugars
sarah.shugars@rutgers.edu
Rutgers University
New Brunswick, USA

Federico Bianchi
fede@stanford.edu
Stanford University
Stanford, USA

Patricia Rossini
patricia.rossini@glasgow.ac.uk
University of Glasgow
Glasgow, United Kingdom

Dirk Hovy
dirk.hovy@unibocconi.it
Bocconi University
Milan, Italy

Rebekah Tromble
rtromble@gwu.edu
George Washington University
Washington, D.C., USA

Nava Tintarev
n.tintarev@maastrichtuniversity.nl
Maastricht University
Maastricht, Netherlands

## ABSTRACT

Increasingly taking place in online spaces, modern political conversations are typically perceived to be unproductively affirming—siloed in so called "echo chambers" of exclusively like-minded discussants. Yet, to date we lack sufficient means to measure viewpoint diversity in conversations. To this end, in this paper, we operationalize two viewpoint metrics proposed for recommender systems and adapt them to the context of social media conversations. This is the first study to apply these two metrics *(Representation and Fragmentation)* to real world data and to consider the implications for online conversations specifically. We apply these measures to two topics—daylight savings time (DST), which serves as a control, and the more politically polarized topic of immigration. We find that the diversity scores for both Fragmentation and Representation are lower for immigration than for DST. Further, we find that while pro-immigrant views receive consistent pushback on the platform, anti-immigrant views largely operate within echo chambers. We observe less severe yet similar patterns for DST. Taken together, Representation and Fragmentation paint a meaningful and important new picture of viewpoint diversity.

## CCS CONCEPTS

• **Computing methodologies → Discourse, dialogue and pragmatics**.

## KEYWORDS

Twitter, conversation network, viewpoint diversity, echo chambers

*WSDM '23, February 27-March 3, 2023, Singapore, Singapore*
© 2023 Association for Computing Machinery.
ACM ISBN 978-1-4503-9407-9/23/02…$15.00
https://doi.org/10.1145/3539597.3570487

**ACM Reference Format:**
Rishav Hada, Amir Ebrahimi Fard, Sarah Shugars, Federico Bianchi, Patricia Rossini, Dirk Hovy, Rebekah Tromble, and Nava Tintarev. 2023. Beyond Digital "Echo Chambers": The Role of Viewpoint Diversity in Political Discussion. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining (WSDM '23), February 27-March 3, 2023, Singapore, Singapore.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3539597.3570487

## 1 INTRODUCTION

Political conversations between everyday people form the foundation of a healthy democracy [34]. In theory, exchanging perspectives allows citizens to collaboratively identify the best solutions to shared problems and builds democratic legitimacy for the implementation of those solutions [18, 20]. In practice, however, there are many reasons to be skeptical that these political conversations are actually achieving their goals. Increasingly taking place in online spaces, modern political conversations are typically perceived to be unproductively affirming—siloed in so called "echo chambers" of exclusively like-minded discussants [4, 6]. However, this focus on the worldview or ideology of discussants overlooks a crucial ingredient of discursive democratic theory: the viewpoints expressed in conversation. For democratic discourse to be productive, it is in some ways less important that the interlocutors themselves embrace different ideologies than that they are aware of and come into contact with different views [36]. On relatively public and open social media platforms, such as Reddit or Twitter, it may be more likely for those who hold a particular opinion to encounter divergent views as part of comment and reply threads. Yet to date, most research examining echo chambers has focused on the ideology of users and those they follow, rather than the specific viewpoints they engage with or to which they are exposed [4–7, 16].

The largely understudied dimension of viewpoint diversity serves as the primary focus of this work. We take inspiration from the viewpoint diversity metrics conceptualized for the news recommender systems domain by Vrijenhoek et al. [48]. We adapt the definition and propose novel operationalization of two such metrics to measure viewpoint diversity for the domain of social media conversations. This is the first study to apply the two metrics to real world data, in this case, online conversations. We also present in-depth analysis of the metric behavior and discuss what it means in the context of deliberative democratic theory.

*Representation* is a conversation-level measure which measures how the views expressed in a single conversation compare to the breadth of views expressed overall. This measure allows us to assess the overall prevalence and distribution of various oppositional vs. supportive viewpoints across conversations. However, it does not capture whether individual participants directly engage with alternative viewpoints. Nor does it tell us much about the exchange of viewpoints within any given conversation. *Fragmentation* in contrast, is a user-level metric which allows us to assess whether and how viewpoints are placed in dialogue with one another, and as such, gives richer meaning and content to the analysis of echo chambers. In contrast to previous work, we consider the specific viewpoints user engage with or to which they are exposed to, rather than the worldview or ideology of discussants. We apply these measures to two topics—daylight savings time (DST), which serves as a control, and the more political topic of immigration. We have selected these two topics because they allow us to situate our measures of Representation and Fragmentation within the larger, platform-level context.

On Twitter, which is known for irreverence and overall negativity [45], we might expect oppositional claims to dominate on virtually any given topic. Yet, from a democratic perspective, such oppositional stance-taking is not inherently problematic. That people complain a lot about an issue like DST does not spell doom for democracy. Nor would democracy be in jeopardy if those who complain about DST rarely encounter pro-DST views. If, however, we see a more extreme imbalance for a salient political topic such as immigration, then we do have reason for concern.

We find that Twitter conversations contain relatively few viewpoints overall (i.e., most tweets are observational or informational, not stance-taking), when users do express viewpoints on immigration in the U.S., anti-immigration views dominate. This tendency towards oppositional immigration viewpoints is even more extreme than negativity towards DST—suggesting that this is more than a mere reflection of platform culture. Our further findings are troubling: anti-immigrant views are rarely countered by pro-immigrant views. Where viewpoint interactions do occur, it is largely because pro-immigration views receive anti-immigrant replies. In other words, while pro-immigrant views receive consistent pushback on the platform, anti-immigrant views largely operate within echo chambers. As discussed in greater detail below, these findings are particularly concerning in light of previous research on the role of such echo chambers in generating attitude extremity [8, 11, 44], spirals of silence [14, 26], and asymmetric polarization [21, 29].

Overall, this work highlights the importance of examining viewpoint diversity, not just ideological diversity. Our measures of Representation and Fragmentation provide tools for examining viewpoint interactions at both the conversational and individual level, and in turn provide important insights into the democratic health of online conversations. By comparing the salient political topic of immigration to the control topic of DST, we not only demonstrate the presence of echo chambers on Twitter, but illustrate how this effect is more extreme for political conversations. [1]

This work is a collaborative effort of researchers from computational, social, and political backgrounds. With this paper we also want to emphasize the importance of interdisciplinary research to have a well rounded understanding of the problem. This helps in coming up with holistic solutions and interventions that are beyond the capabilities of general machine learning (ML) models which are made from a very computational perspective. Social and political scientists in the team formulated the nuanced labeling of the data, while computer scientists complemented their efforts to build predictive models and derive insights from the data. Only together could we situate the insights in the context of political discourse and what it means for a deliberative democratic society. This highlights the importance of building other tools and applications that can leverage high quality data, instead of just focusing on building yet another ML model.

## 2 RELATED WORK

From a normative perspective, we draw heavily on work in deliberative democracy, which argues that political conversations between citizens form the foundation of democratic life [18, 20, 25, 34]. This literature has examined conversation health in various settings [35, 47], but the focus on conversational dynamics has gained renewed attention in light of the forms of interpersonal, group, and mass communication enabled by social media. Measures of the quality of conversations (e.g., toxicity, rationality, and mutual respect) [35] have perhaps received the most attention, but homophily and echo chambers have also proven important in the literature [2, 32].

However, most work on online echo chambers focuses on users' networks [4–7, 16] the content to which users may be exposed—for example, analyzing the political alignment of news outlets based on URLs contained in a post [4, 22], or the news media accounts users follow [1]. Little previous research examines the specific claims or viewpoints that users encounter [13, 41].

While previous studies suggest that people tend to follow user and organizational accounts with similar political leanings [1], post content from ideologically uniform sources [22], and encounter and engage with content from ideologically-aligned news sites [4, 22], this body of work is unable to tell us whether and to what extent users engage with divergent viewpoints within and across social media conversations.

Recent literature in the news recommender domain has drawn inspiration from the "Democratic Notions of Diversity" that focuses on grand concepts such as democracy, freedom of speech, inclusion, mutual respect and tolerance [27, 33, 48]. In their work, Helberger[27] describes four most commonly used theories discussing the democratic role of media: Liberal, Participatory, Deliberative, and Critical model. As mentioned and described earlier we draw heavily on work in deliberative democracy. Vrijenhoek et. al. [48] propose 5 metrics, adapted from existing Information Retrieval practices to measure viewpoint diversity of ranked lists of recommendations by news recommenders and quantitatively evaluate the various democratic notions of diversity. Inspired from

[1] Code and sample data available at https://github.com/hadarishav/beyond-digital-echo-chambers

their conceptualization of Representation and Fragmentation we adapt and operationalize the two metrics for social media conversations. We apply the two metrics on Twitter conversations, present an in-depth analysis of the metric behavior and what it means in the context of deliberative democratic theory and echo-chambers.

This study extends our understanding of political democratic discourse generally, and echo chambers specifically, by studying these viewpoint-based dynamics. One significant computational challenge to viewpoint-based analysis is identifying what viewpoints are present in a conversation. The closest work focuses primarily on natural language inference (NLI) or stance detection (e.g., [30]). In NLI, for given pairs of sentences (premise and hypothesis), the task is to predict whether the hypothesis given is True, False or not related with respect to the premise. In stance detection a text is labeled as being for, against, or neutral towards some target topic. Some recent work has also focused on determining the strength of stance and the logic of evaluation that reflects the general perspective behind the stance [19]. While these approaches are increasingly more accurate in capturing stance, the label inferred by stance detection does not necessarily reflect what we typically mean by “opinion” or viewpoint [28, 42]. In line with democratic discourse theory, we are primarily interested in whether or not a *claim* is made within a text [47]. This is a more subtle notion than stance and implies the presence of an argument, not just an opinion. Here, we therefore develop our own claim detection classifier, as discussed in Section 3.3.

While Reply Trees [15, 17, 24, 31, 38, 43, 49] are by far the most common way to model conversation networks, the literature has taken a range of other approaches, such as Mention Graphs [17], User Graphs [17], and Conversation Cascade [3, 12]. This past work on conversational structure emphasizes the need for both conversation- and individual-level measures. Online conversations take a range of forms, and individuals may have highly divergent experiences based on this overall conversational typology. Therefore, in developing our novel, viewpoint-based approach to examining political discourse, we create two complementary measures of viewpoint diversity—Representation and Fragmentation—which can be meaningfully interpreted across conversational structures.

# 3 DATA COLLECTION AND CLASSIFICATION

As outlined in Figure 1, our pipeline for data collection and annotation consists of several steps at both the tweet and conversation level. Each of these steps is described in detail below.

## 3.1 Tweet Collection

For both topics, daylight savings time (DST) and immigration, we used a keyword-based approach to identify relevant tweets. Keywords were selected by social scientists in a multi-stage process. We restricted both samples to English-language tweets and used the Enterprise streaming API in early 2020 to collect tweets in real time. Due to the time-sensitive nature of DST, the data collection for this topic was conducted between $8^{th}$ and $10^{th}$ of March 2020.

## 3.2 Tweet Annotation

**Topical relevance annotation.** At this stage in our pipeline, we were interested in retaining only those tweets that were relevant for our topics. We therefore developed a relevance classifier for each. These were trained on annotations of 9,814 tweets on DST and 9,931 tweets on immigration respectively. A team of five trained undergraduate annotators from George Washington University manually labeled a random sample of tweets collected based on our keywords. Annotators were told which topic the tweet was collected for and were asked to classify the tweet as either relevant, irrelevant, or “not English.”

The relevance classifier trained on these annotations is described in further detail in Section 3.3.

**Viewpoint annotation.** Using the relevance classifier, we identify a final seed corpus of 10,529 DST tweets and 15,119 immigration tweets that were then annotated by the same group of undergraduate students for the presence or absence of **diagnostic claims** and **counter-claims**. Following previous work [46, 47], *diagnostic claims* highlight a problem associated with a topic and represent an *oppositional viewpoint* in relation to that topic. For example, a tweet about daylight savings time might contain a diagnostic claim that suggests that DST interrupts sleep schedules. For the *immigration* topic, *diagnostic claims* identified a problem with immigrants or permissive immigration policies. In other words, they were explicitly anti-immigrant/anti-immigration. In contrast to diagnostic claims, *counterclaims* counter the concerns identified in the former. They are considered “counterclaims” even when made in isolation from the diagnostic claims they seek to counter. For example, a single tweet that describes how DST helps sleep schedules would be considered to include a counterclaim in our framework (i.e., it logically counters the problem diagnosis that DST interrupts sleep, even if that diagnosis is not explicitly made), as would a tweet that suggests that immigrants benefit the economy (i.e., since it logically counters the problem diagnosis that immigrants harm the economy).

Pairs of students independently annotated tweets in batches (mean batch size = 302 tweets). In order to prevent discrepancies developing across annotators, the student pairs rotated with each batch, with the fifth student in each rotation attending the resolution meetings, observing and sharing any apparent discrepancies with the full team. The team then collectively agreed on a standard and clarified the annotation guidelines accordingly. The period in which these standards were being set and updated in the guidelines was treated as a training phase.

Once the team annotated three batches without any changes to the guidelines *and* inter-annotator agreement was consistently high (above 0.80 for percent agreement and above 0.70 for Krippendorff’s alpha), the team began full annotation. The procedures remained the same during this phase, with one student continuing to observe resolution meetings, but no guideline updates were deemed necessary. We measured pairwise inter-annotator agreement based on whether both annotators agreed that a viewpoint was or was not present. For the DST dataset, percent agreement was 0.87 and Krippendorff’s alpha was 0.72. For the immigration dataset, percent agreement was 0.85 and Krippendorff’s alphas was 0.71.

The dataset is still under development i.e. we are adding more fine grained labels. This full dataset will be published in a separate paper with further details about its curation. For this paper, we release a sample of the data for the purposes of reproducibility. We see the main contribution of this work as the operationalization and

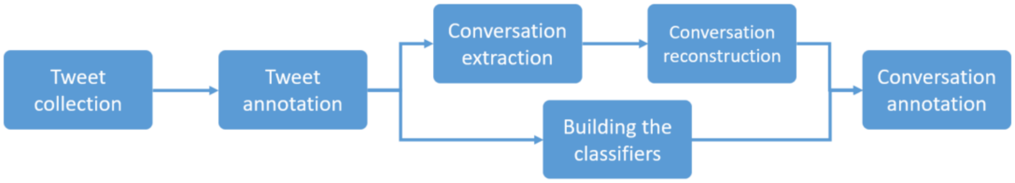


**Figure 1: Our pipeline for data collection and classification.**

interpretation of our metrics of viewpoint diversity in the context of social media conversations, described in the subsequent sections.

## 3.3 Relevance and Viewpoint Detection Models

We made use of recent neural language models [39, 40] to build separate classifiers to predict the (1) relevance of the tweets to our topics and (2) to determine whether the tweets contain any diagnostic claims, counterclaims, or no viewpoint. We built these classifiers separately for each topic, resulting in a total of four trained classifiers.

We used BERTweet [37] a large language model pre-trained on tweets and fine-tuned it with our datasets (BERTweet is used as encoder to which we add an additional classification layer). For each topic and classification task, we trained a model for 4 epochs on 80% of the data, with a batch size of 32. We used 10% of the data to evaluate it every 20 steps. Finally, we picked the model with the lowest validation loss and we evaluate it on the last 10% of the data.

Overall, our models return relatively accurate results across tasks and topics, including the hardest classification tasks. The **relevance model** for *DST* had a macro-F1 score of .95 , with a precision of .92 and a recall of .98. Similarly, the trained relevance model for *immigration* had a macro-F1 score of .92, identifying relevant tweets with a precision of .92 and a recall of .92. **Viewpoint classification** proved to be a harder task, but our models achieved a macro-F1 of .80 for *DST* and a macro-F1 of .80 for *immigration*. For both models, identification of counterclaims was particularly challenging, with this class achieving a precision of .75 and a recall of .78 for DST, and a precision of .70 and a recall of .74 for immigration.

## 3.4 Conversation extraction and reconstruction

**Representing Conversations.** After the annotation of our seed tweets, we collected and reconstructed the conversations of which the seed tweets are a part. To do this, we first operationalized the notion of a "conversation". Of the four approaches summarized in the related work, we follow the model of reply trees. [2] This choice is suitable for examining the degree to which conflicting views come into contact with each other, since they show exactly what content is shared *in response* to other content. More formally, reply trees are directed, acyclic graphs with at least two nodes. Each node represents a single tweet, and each edge shows a reply from a newer tweet to an older tweet (c.f., Figure 2).

Any given tweet may receive any number of replies, but may only be made in response to at most one tweet. Hence the resulting graph is strictly acyclic and contains no loops. The root node *A* in Figure 2 has an out-degree of zero, indicating that it is not in response to any other tweet. All other tweets have an out-degree of 1, indicating that they have exactly one parent to which they are responding. Furthermore, the nodes can have any in-degree value–e.g., can receive any number of replies. A tweet which receives no replies (in-degree of zero), is defined as a leaf. Note also that each tweet is associated with a user, and each user may author any number of tweets.

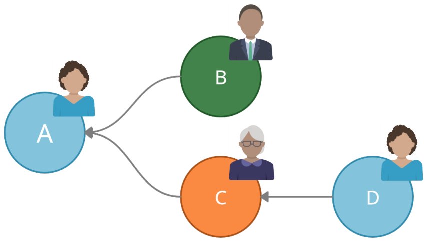


**Figure 2: A conversation network containing four tweets of A, B and C, and D. The initial tweet A is the root node of the tree. Tweets B and C are both replies to A, while D is a reply to tweet C. Tweets A and D are written by the same author.**

**Conversation Extraction.** We used the Academic Research Track of the Twitter API to reconstruct conversations. Under this new track of the API, each tweet is associated with a *conversation_id* which is shared across all tweets in a unique reply tree. This approach allows us to make calls for each conversation.

Based on initial API calls using both DST and immigration seed tweets, we decided to retrieve a maximum of 50 tweets per *conversation_id*. This limit allows us to capture the majority of tweets connected to each conversation while keeping the computational task tractable. We further defined a conversation to require at least two different tweets from at least two different authors. For the purposes of this analysis we discard any singleton tweets which received no replies, as well as any single-author threads.

**Conversation Reconstruction.** After extracting the conversation tweets, we reconstructed the conversation tree by using the *referenced_tweets* information from each retrieved tweet object. For each tweet, we examined whether it had a *referenced_tweets* of type *replied_to*

If this value is None, the given tweet is a root node and has out-degree of zero. Otherwise, this field lists the unique *id* of that tweet's parent—the single tweet to which the examined tweet is replying. We then reconstructed each reply tree by iteratively attaching each child to its parent. For example, we connect tweet A to tweet B, if

[2] We consider only replies. We do not consider other forms of engagement like quote tweet and retweet.

*referenced_tweets* → *id* field in the former is equal to the *id* field in the latter. [3]

**Conversation dataset.** Our seed corpus of 10,531 DST tweets is ultimately associated with 1,756 unique conversations. On average, conversations about immigration tended to be longer than those about DST. Our 9,667 seed immigration tweets were associated with 404 unique conversations. [4]

### 3.5 Conversation annotation

Given reconstructed conversations, the final step was to identify the relevance and viewpoint status of each tweet in our conversational corpus. Table 1 provides a summary of the structural features of our final corpus of conversations, along with the distribution of annotation labels for each topical dataset. The first step for obtaining the labels was to pass every tweet through its topic’s relevance model, which labeled it as either relevant, not relevant or not English (Section 3.3).

Next, we passed tweets through our viewpoint detection models that marked tweets as containing a diagnostic claim, a counterclaim, or none. The relevance and claim detection models are described in Section 3.3. We then aggregated our classification output into four distinct labels. Tweets that are “irrelevant" (L1). Tweets that are relevant but have no diagnostic claim or counterclaim are labeled as “no viewpoint" (L2). Tweets that are identified as including a diagnostic claim are labeled as “Diagnostic claim" (L3) regardless of the output of our relevancy classifier. [5] Similarly, all tweets that are identified as containing a “counterclaim” are labeled as (L4).

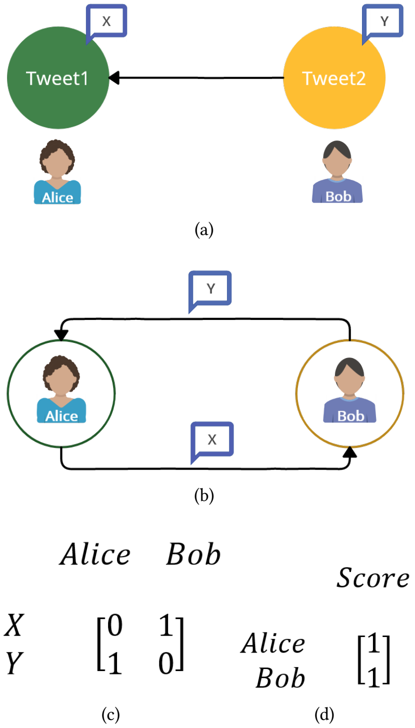
## 4 VIEWPOINT DIVERSITY MEASURES

To better evaluate viewpoint diversity, we introduce a conversation-level measure of Representation and a user-level measure of Fragmentation. Each measure provides insight into the degree to which conflicting views come into contact with each other, either within a conversation or for individual participants. In Section 5, we present the results of these metrics for our DST and immigration datasets.

### 4.1 Fragmentation Diversity

*Definition.* We define the user-level metric of Fragmentation as the complement of the overlap between users’ viewpoint [48]. This metric aims to capture the extent to which individuals within a conversation are exposed to different viewpoints. A value of 1 means people are exposed to maximally different viewpoints, while 0 means people are exposed to the same viewpoints.

*Operationalization.* We define exposure at the level of pair-wise (dyadic) interactions, considering a user to be exposed to a viewpoint if they are replying to a certain viewpoint (e.g., X), or if they receive a reply with a certain viewpoint (e.g., Y). [6] Figure 3(a) illustrates an imaginary conversation in which Alice posts Tweet1 containing viewpoint X and then Bob replies with Tweet2 containing viewpoint Y. In this scenario, Bob is exposed to Alice’s viewpoint through Tweet1 and, after he posts Tweet2, Alice is exposed to Y. If a third user responded to Bob with viewpoint Z, Bob would be exposed to Z, but Alice would not.



(a)

(b)

(c)

(d)

**Figure 3: Measuring the Fragmentation score in an exemplar conversation: (a) The exposure of different viewpoints in a conversation. In this conversation Bob is exposed to the viewpoint X and Alice is exposed to the viewpoint Y, (b) The corresponding viewpoint network to the conversation between Alice and Bob, (c) The viewpoint matrix corresponding to the viewpoint network, and (d) the Fragmentation score for Alice and Bob.**

To measure the degree of overlap in viewpoint exposure, we transform our conversation network into a *viewpoint network*. A viewpoint network is a multi-directed graph in which each node represents a unique user and each edge shows who is exposed to whom and with what viewpoint. In this network, in-degree represents viewpoints a user is exposed to, while out-degree indicates the views the user is disseminating. Figure 3(b) demonstrates the corresponding viewpoint network to the conversation network between Alice and Bob.

We next construct the *matrix representation* of this viewpoint network, as shown in Figure 3(c). Every column in this viewpoint matrix represents a single user, while each row indicates a single viewpoint. Here, we use the viewpoint labels of **L1** (irrelevant), **L2** (no viewpoint), **L3** (diagnostic claim), and **L4** (counterclaim) consistently across all conversations. A conversation with U unique

[3] Each retrieved tweet object includes information on both the retrieved tweet itself and its parent, if there is one. This means that for some of our conversations, the final size of the retrieved conversation was larger than the size limit of 50 that we set.

[4] ≈ 3% and ≈ 14% of the DST and immigration conversations respectively were greater than our set threshold of 50 tweets. We retain 50 tweets each for such conversations.

[5] Certain tweets might be irrelevant but still contain a claim in the context of a conversation.

[6] We do not consider the user’s own viewpoint since we want to observe what other viewpoints a user is exposed to or engages with in a conversation.

| Topic | Structural Information | | | | Annotations | | | |
|---|---|---|---|---|---|---|---|---|
| | # of conversations | # of nodes | # of edges | # of distinct users | Irrelevant (L1) | No viewpoint (L2) | Diagnostic Claim (L3) | Counter Claim (L4) |
| DST | 1756 | 15362 | 13606 | 10578 | 86.85% | 6% | 4.04% | 3.1% |
| Immigration | 404 | 13304 | 12900 | 8611 | 78.43% | 9.86% | 7.7% | 4.01% |

**Table 1: Structural information of conversations and overall distribution of labels in both topics**

authors will therefore have a viewpoint matrix of size 4 x U. This is a positive, weighted matrix in which a user may be exposed to a viewpoint any number of times. Each of these column vectors describe a single user's position in a shared viewpoint space. Similarity between vectors then reflects similarity between the viewpoints that the corresponding users are exposed to. Therefore, to compute the Fragmentation score, we first calculate the similarity between every pair of user vectors (i.e., columns) in the viewpoint matrix.

We calculate this using cosine similarity which ranges between 0 (no overlap) to 1 (complete overlap). For each user in a given conversation, this results in a list of U-1 similarity scores. Next, we take the mean of each user's pairwise similarity scores, and finally subtract this value from 1 since Fragmentation and overlap (similarity) have an inverse relationship. Recall, a Fragmentation value closer to 1 means that the user is exposed to maximally different viewpoints from their conversational peers, and Fragmentation value closer to 0 means that the user is exposed to maximally similar viewpoints from their conversational peers.

## 4.2 Representation Diversity

*Definition.* For our Representation metric, we adapt the definition and operationalization from [48] for the context of social media conversations. Representation compares the views expressed in a single conversation to the breadth of views expressed for the topic overall. Representation thereby provides a *conversation-level metric* which denotes the degree to which conversations are restricted to certain views or capture the diversity of possible opinions. A Representation score of 0 indicates that the distribution of viewpoints in a conversation is similar to the overall distribution of viewpoints in the topical data pool. As we move towards a Representation score of 1 the discrepancies between a given conversation and the data pool increases, while scores closer to 0 indicate that the conversation is "typical" or representative in terms of viewpoint diversity (in the context of a given topic).

*Operationalization.* We compute Representation by measuring the Kullback-Leibler (KL) divergence between the probability distribution of the viewpoint categories in a single conversation to the viewpoint distribution in the entire pool of conversations data for a given topic. Here, the possible viewpoint categories again refers to the labeling system described in Section 3.2. We measure the KL divergence between the two distributions and then normalize the value for each conversation with the maximum KL divergence value obtained.

# 5 RESULTS AND ANALYSIS

In this section, we report our results on the Fragmentation and Representation metrics for Twitter conversations for two topics: immigration and daylight savings.

## 5.1 Fragmentation Diversity

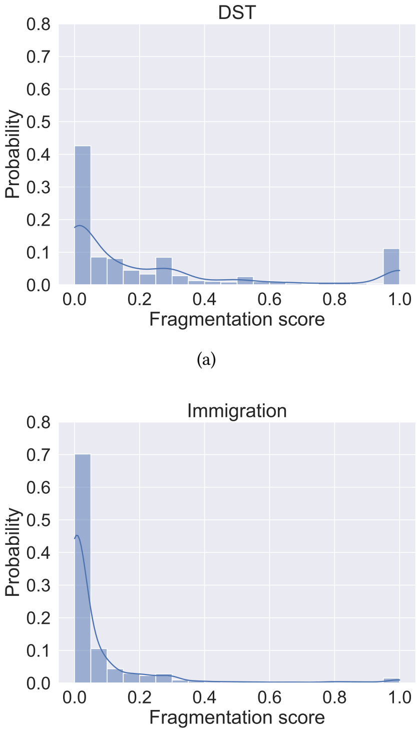
Figure 4 shows the distribution of Fragmentation values for both topics. As described in Section 4, Fragmentation is the complement of overlap between individuals' viewpoint exposure in a conversation. Recall that a user with a Fragmentation score closer to 0 is exposed to the same viewpoints as their peers.[7] In DST conversations, more than 40% of users have Fragmentation scores between 0—0.05, indicating that users in these conversations have a high overlap among the exposed viewpoints. In contrast, a user with a Fragmentation score closer to 1 is exposed to viewpoints different from their conversational peers. On the other hand, we also see that over 10% of users in these conversations have a Fragmentation score of over 0.95, indicating that a notable number of users do often diverge from the majority viewpoints being discussed in the conversations. Interestingly, this finding is even more extreme within the conversations about immigration. Nearly 70% of users have a Fragmentation score between 0—0.05, and virtually none have a Fragmentation score near 1.

It is not evident whether the prevalence of L1 ("Irrelevant") viewpoints in our dataset should be considered noise. Therefore, we also compute our results of Fragmentation without considering L1 viewpoints. That is, we did not use the L1 values from the viewpoint matrix when calculating similarity between users. We notice a similar distribution as before for Fragmentation values for both topics without L1 viewpoints.

## 5.2 Representation Diversity

Representation is a conversation-level diversity metric that compares the views expressed in a single conversation to the breadth of views expressed for the topic overall. Figure 5 shows the distribution of Representation scores for both topics. Recall that a score close to 0 indicates that a conversation's distribution of viewpoints matches the distribution in the overall pool of viewpoints for the topic. We see that over 20% of DST conversations have a low level of diversity (score between 0—0.05), whereas over 60% of conversations about immigration fall into that bin. This suggests that the majority of individual immigration conversations mirror the viewpoint distribution compared to all conversations on immigration (described in Table 1). In contrast, individual conversations about

[7] Note that that even if a user is exposed to exactly one viewpoint, their Fragmentation score could be 0, if their peers are exposed to exactly one viewpoint as well (a limitation addressed by the metric of Representation).



(a)

(b)

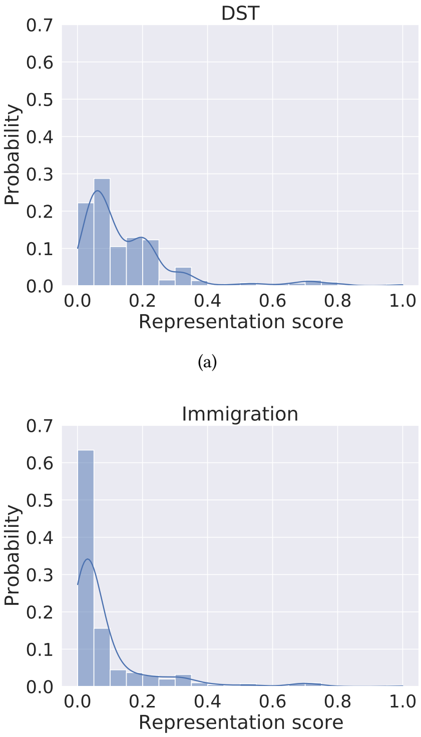
**Figure 4: The distribution of Fragmentation for daylight savings time (DST) (a) and immigration (b) conversations. The x-axis shows a Fragmentation score per user. (*Fragmentation is defined as the complement of the overlap between users’ viewpoint. A high fragmentation value means people are exposed to maximally different viewpoints, while a low value means people are exposed to the same viewpoints in a conversation.*)**

DST tend to have more variability from the full set of conversations on DST. However, it is also interesting to note that Representation scores for immigration also have a much longer tail—indicating that while these conversations are most likely to to be similar to the pool distribution, some of the deviations are also extreme.

Again, we consider the influence of the L1 viewpoint (irrelevant) on the metric performance. Similar to Fragmentation, we compute Representation results without considering L1 viewpoints. Operationally this means that we did not consider the L1 values when computing the KL divergence between the distribution of the pool and the conversation. We again observed similar distributions as before for Representation values for both topics without L1 viewpoints.

## 5.3 Dyadic Interactions

To investigate how the stance-taking viewpoints i.e., oppositional claim (L3) and supporting counterclaim (L4) engage with each other in a conversation we further examine their pair-wise (dyadic) interactions. To focus on substantive conversation, we ignore all other dyadic interactions (with L1 and L2). We calculate the conditional probability of $P(L_i|L_j)$ where $L_i$ and $L_j$ represents labels with $i, j \in \{3, 4\}$. It represents the likelihood of a reply with label $L_i$ to a tweet with label $L_j$. For instance, $P(L_4|L_3) = 0.33$ means that, given a tweet with an oppositional claim, the likelihood that a replying tweet has a supporting counterclaim is 0.33.



(a)

(b)

**Figure 5: The distribution of Representation for daylight savings time (DST) (a) and immigration (b) conversations. The x-axis shows a Representation score per conversation. (*Representation compares the views expressed in a single conversation to the breadth of views expressed for the topic overall. As we move from lower to higher value the divergence between the distribution of viewpoints in the conversation and the topical data pool increases.*)**

Here, we see evidence for echo-chambers occurring, particularly among users expressing the oppositional (L3) viewpoints for immigration. For the control topic of DST, we see that when beginning with tweets that contain the oppositional viewpoint (L3)—a responding tweet has a 62% chance of similarly containing the oppositional viewpoint. For immigration, that probability rises to 77%. In other words, more than three-quarters of responses to anti-immigrant claims also include anti-immigrant viewpoints. Those who voice support for immigrants (L4) in contrast have a lower likelihood of receiving a L4 reply (49%). In other words, we can see this as more echo-chamber interactions on the oppositional side for immigration.

What is perhaps more worrying is that those who voice support for immigrants (L4) also have a 51% chance of receiving an anti-immigrant reply (L3). Again, we see that this effect is even more extreme for our political issue than it is for DST. A user who supports DST (L4) has only a 38% chance of receiving an anti-DST (L3) reply.

Compared to conversations about DST, we see that conversations espousing anti-immigrant viewpoints tend to take place in echo chambers, while comments supporting immigrants are frequently met by anti-immigrant retorts. While the difference in response may not be dramatic across the two topics, it is concerning that opposition to immigration appears to go largely unchallenged, while support for immigration receives regular pushback.

## 6 CONCLUSION

In this section, we discuss the broader implications of our findings, limitations of our data and methods, possible directions for future work, and some additional applications of the measures developed.

*Broader Implications.* Bringing these findings together, our work suggests that oppositional viewpoints, such as those opposing daylight savings or immigrants and immigration, are more common in Twitter conversations than supportive viewpoints and are more likely to exist within echo chambers. While this can be partially attributed to the overall culture of the platform under study, we see that this effect is even more pronounced for immigration than for the less politically salient topic of DST. Our conversation-level measure of Representation shows that conversations about both immigration and DST reflect the distribution of opinions in the full population.

Our more fine-grained, individual-level measure of Fragmentation further shows that individual users tend to have very little variability in the viewpoints to which they are exposed. This effect is more pronounced for immigration than for DST. Only when looking at these metrics together can we start to diagnose the presence of echo chambers: a low Representation value of most conversations show that they have skewed distribution and consist of the majority (irrelevant or oppositional) viewpoints. While, a low Fragmentation score for most users show that they are exposed primarily to these majority viewpoints.

Our analysis of pair-wise responses also indicates that for the topic of immigration users who share oppositional claims are most likely to receive like-minded replies, while users who make counterclaims receive both like-minded and oppositional replies. This suggests distinctly divergent user experiences—users who make diagnostic claims, may interact primarily in echo chambers while users who make counterclaims in support of the issue at hand may be regularly forced to defend and justify their views. It is particularly troubling that this dynamic is more prevalent within our politically salient topic of immigration (compared to the topic of DST)—suggesting that this pattern may be more pronounced for political talk.

Though this study alone cannot diagnose the extent to which such echo chambers harm democratic discourse on Twitter, previous research suggests that online echo chambers can promote attitude extremity [8, 11, 44]. And the one-sided nature of the echo chambers we find in this study points to further concerns about the silencing of pro-immigrant voices, as fear of (or exhaustion from) backlash may lead to a spiral of silence [14, 26]. And it bolsters larger concerns that *asymmetric* polarization offers a breeding ground for intolerance of and discrimination against marginalized communities [21, 29].

*Limitations and Future Work.* Rate limits of the Twitter API[8] continue to be a major bottleneck for conversational research. This challenge is compounded by deleted tweets or accounts marked as private. This missing data can then lead to disconnected components in conversations, which is a notable limitation for conversational research. For the datasets in this paper, we found that 18% of DST while 63% of immigration conversations resulted in disconnected components.

Firstly, we chose a specific representation of viewpoint labels. In future work, we will expand our datasets by incorporating even more fine-grained labels that allow us to explore the range and diversity of viewpoints *within* diagnostic and counterclaim categories. Secondly, even though our analysis with and without irrelevant viewpoint shows similar patterns, the prevalence of irrelevant tweets in conversations is a considerable limitation when analysing social media conversations. Thirdly, we rely on ML models for our analysis. As is the case when using any automated system, some predictions might be incorrect, and there might be unwanted biases learned by the system. Lastly, the robustness of our results could be further investigated by exploring other design choices and operationalization metrics. There is an imbalance in the number of conversations for DST and immigration. More immigration conversations can be extracted for a more robust comparison. Another direction could be to study other ways of defining a conversation network and how that impacts the viewpoint measures. Future work might also supplement Fragmentation and Representation with other dimensions of viewpoint diversity.

*Additional Applications.* Recent studies have highlighted the importance of carefully curating and documenting datasets on which language models are trained [9, 10, 23]. Bender et al. [10] argue that data dumps taken from the Internet retain only hegemonic viewpoints overrepresenting younger users and those from developed countries. The authors propose datasets should be created with a thoughtful process such that they are diverse in terms of the viewpoints represented. Our measures of Representation and Fragmentation can act as essential dimensions of viewpoint diversity for evaluating conversational datasets. Furthermore, in capturing the diverging viewpoint dynamics of different types of conversations, these measures could potentially be used to help identify particularly contentious topics. Detecting low viewpoint diversity could be especially valuable for identifying both individual accounts and semi-organized efforts to intentionally and regularly provide oppositional replies without ever engaging in good-faith exchange.

## ACKNOWLEDGMENTS

This research was supported by a gift from Twitter, Inc. Twitter exerted no control over the direction or findings of the research.

[8] Twitter Developer Platform. 2022. Academic Research Track. (2022). https://developer.twitter.com/en/products/twitter-api/academic-research. Accessed: 2022-01-19.

## REFERENCES

[1] Jisun An, Daniele Quercia, Meeyoung Cha, Krishna Gummadi, and Jon Crowcroft. 2014. Sharing political news: the balancing act of intimacy and socialization in selective exposure. *EPJ Data Science* 3 (2014), 1–21.

[2] Sinan Aral, Lev Muchnik, and Arun Sundararajan. 2009. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences* 106, 51 (2009), 21544–21549.

[3] Arunkumar Bagavathi, Pedram Bashiri, Shannon Reid, Matthew Phillips, and Siddharth Krishnan. 2019. Examining untempered social media: analyzing cascades of polarized conversations. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 625–632.

[4] Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (2015), 1130–1132.

[5] Pablo Barberá. 2015. Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political analysis* 23, 1 (2015), 76–91.

[6] Pablo Barberá, John T Jost, Jonathan Nagler, Joshua A Tucker, and Richard Bonneau. 2015. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science* 26, 10 (2015), 1531–1542.

[7] Marco Bastos, Dan Mercea, and Andrea Baronchelli. 2018. The geographic embedding of online echo chambers: Evidence from the Brexit campaign. *PloS one* 13, 11 (2018), e0206841.

[8] Fabian Baumann, Philipp Lorenz-Spreen, Igor M Sokolov, and Michele Starnini. 2020. Modeling echo chambers and polarization dynamics in social networks. *Physical Review Letters* 124, 4 (2020), 048301.

[9] Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6 (2018).

[10] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proc. of FAccT '21*.

[11] Yochai Benkler, Robert Faris, and Hal Roberts. 2018. *Network propaganda: Manipulation, disinformation, and radicalization in American politics.* Oxford University Press.

[12] John Bollenbacher, Diogo Pacheco, Pik-Mai Hui, Yong-Yeol Ahn, Alessandro Flammini, and Filippo Menczer. 2021. On the challenges of predicting microscopic dynamics of online conversations. *Applied Network Science* 6, 1 (2021), 1–21.

[13] Iván Cantador, María E Cortés-Cediel, and Miriam Fernández. 2020. Exploiting Open Data to analyze discussion and controversy in online citizen participation. *Information Processing & Management* 57, 5 (2020), 102301.

[14] Hsuan-Ting Chen. 2018. Spiral of silence on social media and the moderating role of disagreement and publicness in the network: Analyzing expressive and withdrawal behaviors. *New Media & Society* 20, 10 (2018), 3917–3936.

[15] Daejin Choi, Jinyoung Han, Taejoong Chung, Yong-Yeol Ahn, Byung-Gon Chun, and Ted Taekyoung Kwon. 2015. Characterizing conversation patterns in reddit: From the perspectives of content properties and user participation behaviors. In *Proceedings of the 2015 acm on conference on online social networks*. 233–243.

[16] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences* 118, 9 (2021).

[17] Peter Cogan, Matthew Andrews, Milan Bradonjic, W Sean Kennedy, Alessandra Sala, and Gabriel Tucci. 2012. Reconstruction and analysis of twitter conversation graphs. In *Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research*. 25–31.

[18] Joshua Cohen. 1989. Deliberation and Democratic Legitimacy. In *The Good Polity: Normative Analysis of the State*, Alan Hamlin and Phillip Petit (Eds.). Blackwell, New York.

[19] Tim Draws, Oana Inel, Nava Tintarev, Christian Baden, and Benjamin Timmermans. 2022. Comprehensive Viewpoint Representations for a Deeper Understanding of User Interactions With Debated Topics. In *ACM SIGIR Conference on Human Information Interaction and Retrieval* (Regensburg, Germany) *(CHIIR '22)*. Association for Computing Machinery, New York, NY, USA, 135–145. https://doi.org/10.1145/3498366.3505812

[20] John S. Dryzek. 2009. Democratization as Deliberative Capacity Building. *Comparative Political Studies* 42, 11 (2009). https://doi.org/10.1177/0010414009332129

[21] Deen Freelon, Alice Marwick, and Daniel Kreiss. 2020. False equivalencies: Online activism from left to right. *Science* 369, 6508 (2020), 1197–1201.

[22] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *Proceedings of the 2018 World Wide Web Conference*. 913–922.

[23] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for Datasets. *CoRR* abs/1803.09010 (2018).

[24] Maria Glenski, Emily Saldanha, and Svitlana Volkova. 2019. Characterizing speed and scale of cryptocurrency discussion spread on reddit. In *The World Wide Web Conference*. 560–570.

[25] Jürgen Habermas. 1984. *The theory of communicative action.* Beacon Press, Boston.

[26] Keith N Hampton, Harrison Rainie, Weixu Lu, Maria Dwyer, Inyoung Shin, and Kristen Purcell. 2014. *Social media and the 'spiral of silence'*. PewResearchCenter.

[27] Natali Helberger. 2019. On the Democratic Role of News Recommenders. *Digital Journalism* 7, 8 (2019), 993–1012. https://doi.org/10.1080/21670811.2019.1623700 arXiv:https://doi.org/10.1080/21670811.2019.1623700

[28] Kenneth Joseph, Sarah Shugars, Ryan Gallagher, Jon Green, Alexi Quintana Mathé, Zijian An, and David Lazer. 2021. (Mis)alignment Between Stance Expressed in Social Media Data and Public Opinion Surveys. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[29] Daniel Kreiss. 2021. "Social Media and Democracy: The State of the Field, Prospects for Reform," edited by Nathaniel Persily and Joshua A. Tucker.

[30] Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys (CSUR)* 53, 1 (2020), 1–37.

[31] Ravi Kumar, Mohammad Mahdian, and Mary McGlohon. 2010. Dynamics of conversations. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. 553–562.

[32] Jae Kook Lee, Jihyang Choi, Cheonsoo Kim, and Yonghwan Kim. 2014. Social media, network heterogeneity, and opinion polarization. *Journal of communication* 64, 4 (2014), 702–722.

[33] Felicia Loecherbach, Judith Moeller, Damian Trilling, and Wouter van Atteveldt. 2020. The Unified Framework of Media Diversity: A Systematic Literature Review. *Digital Journalism* 8, 5 (2020), 605–642. https://doi.org/10.1080/21670811.2020.1764374 arXiv:https://doi.org/10.1080/21670811.2020.1764374

[34] Jane Mansbridge. 1999. Everyday talk in the deliberative system. In *Deliberative Politics: Essays on Democracy and Disagreement*, Stephen Macedo (Ed.). Oxford University Press, 1–211.

[35] Jane Mansbridge. 2015. A minimalist definition of deliberation. *Deliberation and development: Rethinking the role of voice and collective action in unequal societies* (2015), 27–50.

[36] Hugo Mercier and Hélène Landemore. 2012. Reasoning is for arguing: Understanding the successes and failures of deliberation. *Political Psychology* 33, 2 (2012), 243–258.

[37] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 9–14.

[38] Ryosuke Nishi, Taro Takaguchi, Keigo Oka, Takanori Maehara, Masashi Toyoda, Ken-ichi Kawarabayashi, and Naoki Masuda. 2016. Reply trees in twitter: data analysis and branching process models. *Social Network Analysis and Mining* 6, 1 (2016), 26.

[39] Debora Nozza, Federico Bianchi, and Dirk Hovy. 2020. What the [mask]? making sense of language-specific BERT models. *arXiv preprint arXiv:2003.02912* (2020).

[40] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics* 8 (2020), 842–866. https://doi.org/10.1162/tacl_a_00349

[41] Robin Schaefer and Manfred Stede. 2021. Argument Mining on Twitter: A survey. *it-Information Technology* 63, 1 (2021), 45–58.

[42] Indira Sen, Fabian Flöck, and Claudia Wagner. 2020. On the reliability and validity of detecting approval of political actors in tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1413–1426.

[43] Sarah Shugars and Nicholas Beauchamp. 2019. Why keep arguing? Predicting engagement in political conversations online. *Sage Open* 9, 1 (2019), 2158244019828850.

[44] Cass Sunstein and Cass R Sunstein. 2018. *# Republic.* Princeton university press.

[45] Rebekah Tromble. 2018. Thanks for (actually) responding! How citizen demand shapes politicians' interactive practices on Twitter. *New media & society* 20, 2 (2018), 676–697.

[46] Rebekah Tromble and Michael Meffert. 2016. The life and death of frames: Dynamics of media frame duration. *International Journal of Communication* 10 (2016), 23.

[47] Rebekah Tromble and Miriam Wouters. 2015. Are We Talking *with* or *past* One Another? Examining Transnational Political Discourse across Western-Muslim "Divides". *International Studies Quarterly* 59, 2 (2015), 373–386. https://doi.org/10.1111/isqu.12167

[48] Sanne Vrijenhoek, Mesut Kaya, Nadia Metoui, Judith Möller, Daan Odijk, and Natali Helberger. 2021. Recommenders with a mission: assessing diversity in news recommendations. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*. 173–183.

[49] Xingshan Zeng, Jing Li, Lu Wang, Nicholas Beauchamp, Sarah Shugars, and Kam-Fai Wong. 2018. Microblog conversation recommendation via joint modeling of topics and discourse. In *Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 375–385.